DOCUMENT RESUME

ED 291 755 TM 011 040

AUTHOR Feldt, Leonard S.; And Others
TITLE Statistical Tests and Confidence Intervals for Cronbach's Coefficient Alpha. Iowa Testing Programs Occasional Papers Number 33.
INSTITUTION Iowa Testing Programs, Iowa City.
PUB DATE Oct 86
NOTE 27p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Confidence Testing; *Measurement Techniques; Sampling; *Statistical Analysis; *Test Reliability
IDENTIFIERS *Confidence Intervals (Statistics)

ABSTRACT

Rigorous comparison of the reliability coefficients of several tests or measurement procedures requires a sampling theory for the coefficients. This paper summarizes the important aspects of the sampling theory for Cronbach's (1951) coefficient alpha--a widely used internal consistency coefficient. This theory enables researchers to test a specific numerical hypothesis about the population alpha and to obtain confidence intervals for the population coefficient. It also permits researchers to test the hypothesis of equality among several coefficients, either under the condition of independent samples or when the same sample has been used for all measurements. The procedures are illustrated numerically, and the assumption and derivations underlying the theory are discussed. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED291755

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 33 - October 1986

Statistical Tests and Confidence Intervals for Cronbach's Coefficient Alpha

Leonard S. Feldt

David J. Woodruff

Fathi A. Salih

Mayuree Srichai

TM011 040

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Leonard Feldt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

ABSTRACT

Rigorous comparison of the reliability coefficients of several tests or measurement procedures requires a sampling theory for the coefficients. This paper summarizes the important aspects of the sampling theory for Cronbach's (1951) coefficient alpha -- a widely used internal consistency coefficient. This theory enables researchers to test a specific numerical hypothesis about the population alpha and to obtain confidence intervals for the population coefficient. It also permits researchers to test the hypothesis of equality among several coefficients, either under the condition of independent samples or when the same sample has been used for all measurements. The procedures are illustrated numerically, and the assumption and derivations underlying the theory are discussed.

Introduction

When an estimate of the reliability of an educational or psychological instrument is needed and the parallel forms and test-retest approaches are impractical, investigators typically rely on internal consistency coefficients. For cognitive tests and affective scales one of the most commonly used indices is Cronbach's (1951) coefficient alpha. This coefficient is also frequently employed in settings which involve raters or observers (Ebel, 1951). The purpose of this review is to summarize the sampling theory for coefficient alpha and to illustrate the uses of this theory in evaluating reliability data.

The experimental problems for which the sampling theory is needed include the following: 1) to test the hypothesis that coefficient alpha equals a specified value in a given population; 2) to establish a confidence interval for the alpha coefficient; 3) to test the hypothesis of equality for two or more coefficients when the estimates are based on independent samples; (4) to test the hypothesis of equality when the observed coefficients are based on the same sample and hence are dependent; and 5) to obtain an unbiased estimate of the population value of alpha.

A test of a specific hypothesis is called for when a revised measurement procedure is compared to an established, accepted procedure. In most such instances this statistical test would involve a directional alternative--that the new procedure is more reliable than the traditional procedure. In some applications the test might be two-ended, however. Such an alternative might arise when changes in a measurement procedure make

administration more efficient, but might affect reliability either positively or negatively.

Studies in which differences among coefficients are of concern to investigators are not uncommon. Research on alternative methods of measuring a specified trait may well call for a test of the equality of alpha coefficients for the several methods. Evaluation of a training program designed to enhance inter-rater reliability may also demand a test of this null hypothesis. Refinement of a instrument may be assessed, in part, by a comparing the reliabilities of several alternative versions. Oaster (1984), for example, encountered this situation in the refinement of Likert scales.

These problems of inference require the development of a sampling error theory for coefficient alpha. The first steps in this development occurred in the early 1960s when Kristof (1963) and Feldt (1965) independently derived a transformation of the sample alpha coefficient which is proven to be distributed as F. They showed how this result can be used to test hypotheses and generate confidence intervals for a single alpha coefficient.

Techniques for testing the equality of alpha coefficients were developed over the following twenty-year period. The first situation to be considered was that of independent coefficients, that is, coefficients obtained from separate examinee samples. Feldt (1969) derived an F test for the two-coefficient case, and seven years later Hakstian and Whalen (1976) extended the methodology to any number of coefficients. Dependent or related coefficients--reliabilities based on the same sample--posed more

complex statistical problems. Feldt (1980) resolved these problems for two coefficients; Woodruff and Feldt (in press) completed the cycle with a test of equality of m dependent coefficients. In each instance, the control of type I error was verified through computer-based Monte Carlo studies.

The present paper synthesizes this statistical theory for Cronbach's alpha. The principal objective is to make the procedures accessible to researchers and to provide numerical illustrations. For each situation the general outlines of the proofs and derivations are presented.

Inference for a Single Alpha Coefficient

Let X_{jp} denote the score of person p on item j. The test consists of n items or parts and is administered to N subjects. Let Y_p denote the total test score for person p, i.e., $Y_p = \Sigma_{j=1}^{n} X_{jp}$. The usual formula for the sample alpha coefficient, which will be denoted as $\hat{\zeta}$ herein, is

$$\hat{\zeta} = \frac{n}{n-1}\left[\frac{\hat{\sigma}_Y^2 - \Sigma_{j=1}^{n}\hat{\sigma}_{X_j}^2}{\hat{\sigma}_Y^2}\right]. \qquad (1)$$

In this formula $\hat{\sigma}_{X_j}^2$ represents the unbiased estimate of the variance for item j, and $\hat{\sigma}_Y^2$ that for score Y. The sample alpha coefficient is denoted by $\hat{\zeta}$ and its parameter value by ζ to avoid confusion with the symbolic representation of statistical significance levels, almost universally denoted in the statistical literature by α.

Following Hoyt (1941), an alternate formula for $\hat{\zeta}$ may be derived by considering the responses of the N subjects on the n items as observations

in a two-way subjects-by-items ANOVA with one observation per cell. Within this framework, a formula for $\hat{\zeta}$ is

$$\hat{\zeta} = \frac{MS(S) - MS(SI)}{MS(S)} = 1 - \frac{MS(SI)}{MS(S)} , \qquad (2)$$

where MS(S) denotes the mean square for subjects and MS(SI) denotes the mean square for subjects-by-items interaction. When applied to the setting in which n raters evaluate N subjects, $\hat{\zeta}$ can be used as measure of inter-rater agreement, with differences among rater means not considered measurement error. In such a case, raters substitute for items in equation (2).

Let E denote expected value and in particular let the expected values for MS(S) and MS(SI) be denoted as E[MS(S)] and E[MS(SI)], respectively. The population value of coefficient alpha is defined as

$$\zeta = \frac{E[MS(S)] - E[MS(SI)]}{E[MS(S)]} = 1 - \frac{E[MS(SI)]}{E[MS(S)]} . \qquad (3)$$

Kristof (1963) and Feldt (1965) independently proved that when the usual assumptions for the two-way random effects (type II) ANOVA are met, the following statistic is distributed as F with N - 1 and (n-1)(N-1) degrees of freedom:

$$\frac{1 - \zeta}{1 - \hat{\zeta}} = \frac{MS(S)/E[MS(S)]}{MS(SI)/E[MS(SI)]} . \qquad (4)$$

It may also be shown that if 1) items are treated as a fixed factor in the two-way ANOVA, 2) the usual assumptions for the two-way mixed model (type III) ANOVA are met, and 3) there is no items-by-subjects interaction, the same F distribution holds for the statistic given in equation (4) (Scheffé, 1959, chap. 8).

The proof that $(1-\zeta)/(1-\hat{\zeta})$ is an F variable follows from the fact that under the assumed ANOVA model MS(S)/E[MS(S)] is distributed as a chi square variable divided by its degrees of freedom, N-1. Likewise, MS(SI)/E[MS(SI)] is distributed as a chi square variable divided by its degrees of freedom, (n-1)(N-1). Under the assumed model these chi squares are independent. Therefore, their ratio (equation 4) is distributed as a central F with N-1 and (n-1)(N-1) degrees of freedom.

This distribution theory for $(1-\zeta)/(1-\hat{\zeta})$ may be used to formulate a test of a specific numerical hypothesis and derive a confidence interval for a population alpha coefficient. To test the null hypothesis $H_o: \zeta = \zeta_o$ against a two-tailed alternative at the α level of significance, let $F(\alpha/2)$ denote the $100\alpha/2$ percentile and $F(1-\alpha/2)$ the $100(1-\alpha/2)$ percentile of the central F with N - 1 and (n-1)(N-1) as its df. The null hypothesis is rejected if

$$\hat{\zeta} < 1 - \frac{(1-\zeta_o)}{F(\alpha/2)} \quad \text{or} \quad \hat{\zeta} > 1 - \frac{(1-\zeta_o)}{F(1-\alpha/2)} \ . \qquad (5)$$

If a one-tailed test at the α level of significance is desired, $\alpha/2$ is replaced by α in the appropriate critical value.

The upper and lower endpoints of a $100(1-\alpha)$ percent interval for for ζ are given respectively by

$$\zeta_U = 1 - [(1-\hat{\zeta})F(\alpha/2)] \quad \text{and} \quad \zeta_L = 1 - [(1-\hat{\zeta})F(1-\alpha/2). \qquad (6)$$

If a one-sided $100(1-\alpha)$ percent interval is desired, $\alpha/2$ is replaced by by α in the appropriate endpoint.

The foregoing results may be illustrated by the following example. Suppose a researcher used 41 examinees to obtain an estimate of .790 for

the alpha coefficient of a 26-item test. The relevant F distribution has df = 40 and 1000, for which the fifth and ninety-fifth percentiles are 0.66 and 1.41. The 90% confidence interval (bounded below and above) has

$$\zeta_L = 1 - (1-.79)(1.41) = .704$$

$$\zeta_U = 1 - (1-.79)(0.66) = .861 .$$

A one-tailed test of H_o: $\zeta_o = .70$, with H_{alt}: $\zeta_o > .70$ and $\alpha = .05$, would require only a lower bound for the critical region. By equation (5), the critical region (C.R.) is

$$\text{C.R.} > 1 - \frac{(1-.70)}{F(.95)} = 1 - \frac{(1-.70)}{1.41} = .787 .$$

Since the observed coefficient alpha of .790 exceeds the lower bound of the critical region, $\zeta_o = .70$ may be rejected.

The expected value of $\hat{\zeta}$, $E[\hat{\zeta}]$, and the bias in $\hat{\zeta}$ can be deduced from the fact that $(1-\hat{\zeta})/(1-\zeta)$ is also distributed as F, but with $(n-1)(N-1)$ and $(N-1)$ degrees of freedom. Since the expected value of a central F is $\nu_2/(\nu_2-2)$, where ν_2 is the second df value,

$$E[(1-\hat{\zeta})/(1-\zeta)] = (N-1)/(N-3),$$

and hence

$$E[\hat{\zeta}] = 1 - (1-\zeta)(N-1)/(N-3).$$

If follows that

$$E[\hat{\zeta}] - \zeta = 2(\zeta-1)/(N-3).$$

Since $\zeta < 1$, the difference must be negative, and hence $\hat{\zeta}$ tends to underestimate ζ. This result was first presented by Kristof (1963).

The negative bias of $\hat{\zeta}$ is generally of little consequence unless N is small. If N = 50 and ζ = .70, for example, the expected value of $\hat{\zeta}$ is .687. With N = 100, the expected value is .694. Where an unbiased estimate of ζ is required, it may be obtained by the formula

$$\tilde{\zeta} = [(N-3)\hat{\zeta}/(N-1)] + 2/(N-1).$$

Comparison of Alpha Coefficients Obtained from Independent Samples

Rigorous comparisons of alternative test scoring procedures, test construction techniques, item formats, item selection strategies, modes of test administration, or competing test instruments entail, in part, the comparison of reliabilities. The first paper to address this problem was published by Feldt (1969), who derived a statistical test of the hypothesis H_o: $\zeta_1 = \zeta_2$. The Feldt approach is based on the test statistic $W = (1-\hat{\zeta}_2)/(1-\hat{\zeta}_1)$. He proved that when the reliability parameters are equal, W is distributed as the product of two independent central F variables. This product, it was shown, could be well approximated by a single F with N_1-1 and N_2-1 degrees of freedom. With modern computing equipment it is relatively simple to determine the probability that a central F will exceed the obtained value of W. If the probability is less than the significance level, the hypothesis of equality can be rejected.

Hakstian and Whalen (1976) extended the methodology to the case of m coefficients. Their test is based on the normalizing transformation of F developed by Paulson (1942) and the fact that $(1-\tilde{\zeta})/(1-\zeta)$ is distributed as F with (N-1)(n-1) and (N-1) degrees of freedom. Paulson proved that

$$z = \frac{(1- \frac{2}{9\nu_2})F^{1/3} - (1- \frac{2}{9\nu_1})}{\sqrt{\frac{2}{9\nu_2}F^{2/3} + \frac{2}{9\nu_1}}},$$

where z is distributed as unit normal deviate. In the present context, the transformation may be stated as follows:

$$z = \frac{(1-\hat{\zeta})^{1/3} - [(1 - \frac{2}{9\nu_1})(1-\zeta)^{1/3} /(1 - \frac{2}{9\nu_2})]}{\sqrt{\frac{18\nu_2(1-\hat{\zeta})^{2/3}}{(9\nu_2 - 2)^2} + \frac{18\nu_2^2(1-\zeta)^{2/3}}{\nu_1(9\nu_2 - 2)^2}}} \quad (7)$$

This ratio implies that $(1-\hat{\zeta})^{1/3}$ is approximately normally distributed with non-zero mean (the term in brackets in equation 7) and variance approximated by

$$S^2 = \left[\frac{18\nu_2(1-\hat{\zeta})^{2/3}}{(9\nu_2-2)^2}\right]\left[1 + \frac{\nu_2}{\nu_1}\right] = \left[\frac{18(N - 1)(1-\hat{\zeta})^{2/3}}{(9N-11)^2}\right]\left[\frac{n}{n-1}\right].$$

Hakstian and Whalen (1976) propose that the weighted average (μ^*) of the $(1-\hat{\zeta}_1)^{1/3}$ be obtained, the weights equalling the reciprocals of the variances. The test statistic is then defined as

$$M = \sum_i^m \left[\frac{(1 - \hat{\zeta}_i)^{1/3} - \mu^*}{S_i}\right]^2, \quad (8)$$

which is interpreted as a chi square with m-1 degrees of freedom. The justification for this interpretation is that the sum of m squared standardized deviations of normal variables from their weighted mean is so distributed. The test is thus analogous to the test of the equality of m correlation coefficients, wherein Fisher's transformation of the coefficients has achieved normality with variances $1/(N_i-3)$. (See Hays, 1981, p. 469.)

There are two minor problems with the Hakstian/Whalen test. First, the variance of $(1-\hat{\zeta}_i)^{1/3}$ is an estimate based on the sample value of ζ_i. This is contrary to theory and in contrast to the case of transformed correlations, in which the variances, $1/(N_i-3)$, do not depend upon sample estimates. Second, even if all ζ_i are equal, the statistics $(1-\hat{\zeta}_i)^{1/3}$ do not come from the same normal distribution unless the bracketed term in equation (7) is the same for all tests. This equality demands that

$$(1 - 2/9\nu_1)/(1-2/9\nu_2)$$

be constant over all tests.

Fortunately, these problems appear to be of little consequence. The use of the sample statistic, $\hat{\zeta}_i$, to replace the parameter, ζ_i, in the second term under the radical in the denominator of equation (7) seems to have little effect on the distribution of the ratio. The net effect might be likened to that of interpreting a t-statistic as a normally distributed variable--an interpretation that involves no serious error when the sample size is larger than 50. (See Marascuilo, 1966) The minimal effect of replacing ζ by $\hat{\zeta}$ in the second term probably results from the fact that this term is of order $2/(9)(N-1)(n-1)$. The first term, which properly includes $\hat{\zeta}$, is of order $2/(9)(N-1)$.

The second problem also proves to be of negligible importance by virtue of the fact that $1-(2/9\nu_1)$ and $1-(2/9\nu_2)$ are both very close to one, regardless of the variation in n_i and N_i from test to test. For example, if $n_1 = 50$ and $N_1 = 100$ the ratio of these terms is 1.0022. If $n_2 = 10$ and $N_2 = 50$, the ratio is 1.0040. Thus, the hypothesis that

$1.0022(1-\zeta_1)^{1/3}$ equals $1.0040(1-\zeta_2)^{1/3}$ is essentially a hypothesis that ζ_1 equals ζ_2.

Woodruff and Feldt (in press) followed a different line of reasoning to arrive at a similar test of the null hypothesis for m coefficients. They adopt the transformation $1/(1-\hat{\zeta})^{1/3}$ rather than $(1-\hat{\zeta})^{1/3}$. A critical point in the subsequent derviation is the identification of a chi-square distribution (df to be determined) for which the variable χ^2/df has nearly the same mean, variance, skewness, and kurtosis as $(1-\zeta)/(1-\hat{\zeta})$. The latter variable is distributed as F with N-1 and (N-1)(n-1) degrees of freedom. The chi-square distribution which best satisfies this requirement takes $df = \tilde{N}_i - 1$, where $\tilde{N}_i = (n_i-1)(N_i)/(n_i+1)$. Woodruff and Feldt approximate the variance of $1/(1-\hat{\zeta})^{1/3}$ by the Wilson/Hilferty (1931) normalizing transformation of a chi-square variable. This leads to the following estimate of the variance of $1/(1-\hat{\zeta}_i)^{1/3}$:

$$S_i^2 = 2/9(\tilde{N}_i-1)(1-\hat{\zeta}_i)^{2/3} \quad .$$

Unlike Hakstian and Whalen (1976), Woodruff and Feldt use the arithmetic mean of the transformed coefficients:

$$\tilde{\mu} = \sum_i^m (1-\hat{\zeta}_i)^{-1/3}/m \quad .$$

Their test statistic, under the assumption of independent samples, is

$$UX_1 = \sum_i^m \left[(1-\hat{\zeta}_i)^{-1/3} - \tilde{\mu}\right]^2 / \bar{S}^2 \quad . \qquad (9)$$

where $\bar{S}^2$ is the arithmetic mean of the several variances, S_i^2. Then, under H_o UX_1 is approximately distributed as χ^2 with m-1 degrees of freedom.

These two approaches may be illustrated by the following data:

Group 1 and Test 1 : $\hat{\zeta}_1 = .784$; $n_1 = 5$, $N_1 = 51$; $(1-\hat{\zeta}_1)^{1/3} = .6$

Group 2 and Test 2 : $\hat{\zeta}_2 = .875$; $n_2 = 5$, $N_2 = 101$; $(1-\hat{\zeta}_2)^{1/3} = .5$

Group 3 and Test 3 : $\hat{\zeta}_3 = .936$; $n_3 = 5$, $N_3 = 151$; $(1-\hat{\zeta}_3)^{1/3} = .4$

The Hakstian and Whalen variances equal .0020179, .00069754, and .00029718. The weighted average, μ^*, equals .4458. The test statistic, interpreted as a χ^2 with df = 2, equals 23.053.

The same data, analyzed via the Woodruff and Feldt approach, yields variances of .0187056, .0134003, and .0139353, and $\overline{\mu} = 2.05556$. The test statistic, also interpreted as a chi-square with df = 2, equals 22.926. One might expect on the basis of the underlying derviations that the Woodruff/Feldt test would give results that are quite consistent with those of the Hakstian/Whalen test, as they were in this instance.

If tests of pairwise contrasts among the coefficients are warranted on the basis of a significant outcome of the omnibus test, the pairs can be considered by Feldt's (1969) test for two coefficients. In the present instance all pairs lead to rejection of the null hypothesis.

Comparison of Alpha Coefficients Obtained from the Same Sample

In some settings it is possible to administer all instruments or apply all procedures to the same sample of N examinees. In such instances the coefficients are statistically dependent, and the test of the null hypothesis must recognize this dependence. To ignore it is tantamount in most applications to adoption of a significance level far more stringent than the nominal level.

The methodology for the case of dependent statistics, like that for independent statistics, was first developed for $H_o : \zeta_1 = \zeta_2$. Feldt (1980) derived three procedures for testing this hypothesis. Simulation studies indicated that all three procedures control type I error rates satisfactorily. Feldt recommended the following test statistic:

$$t = \frac{(\hat{\zeta}_1 - \hat{\zeta}_2)\sqrt{N-2}}{\sqrt{4(1-\hat{\zeta}_1)(1-\hat{\zeta}_2)(1-\hat{\rho}^2)}} \qquad df = N-2 \qquad (10)$$

The squared correlation in the denominator refers to the squared coefficient between the two total-test scores for the sample.

The derivation of this test rests on the fact that if $\zeta_1 = \zeta_2$, then $(1-\hat{\zeta}_2)/(1-\hat{\zeta}_1)$ is distributed identically as the ratio of two dependent sample variances, each with expected value of 1.0. Pitman (1939) proved that the following function of such a ratio is distributed as t with df = N-2:

$$t = \left[(\hat{\sigma}_2^2/\hat{\sigma}_1^2)-1\right]\left[\sqrt{N-2}\right] \Big/ \left[(4\hat{\sigma}_2^2/\hat{\sigma}_1^2)(1-\hat{\rho}^2)\right]^{\frac{1}{2}} .$$

Thus, the same function of $(1-\hat{\zeta}_2)/(1-\hat{\zeta}_1)$ must be distributed as t with df of N - 2. Substitution of $(1-\hat{\zeta}_2)/(1-\hat{\zeta}_1)$ for $\hat{\sigma}_2^2/\hat{\sigma}_1^2$ in this expression untimately leads, after algebraic simplification, to equation 10.

Woodruff and Feldt (in press) extended the methodology to the case of m dependent coefficients. They considered eleven possible test statistics. Extensive Monte Carlo simulation led to three procedures that showed excellent control of type I error and superior power, compared to the others. Of these three technqiues, the procedure identified as UX_1 was the simplest computationally and is summarized here.

As in the case of independent coefficients, Woodruff and Feldt (in press) approximate the variance of $1/(1-\hat{\zeta}_i)^{1/3}$ by the quantity

$$S_i^2 = 2/9(\tilde{N}_i-1)(1-\hat{\zeta}_i)^{2/3} \quad . \tag{11}$$

However, the test for dependent alphas also demands an approximation of the covariance between $1/(1-\hat{\zeta}_i)^{1/3}$ and $1/(1-\hat{\zeta}_j)^{1/3}$. Using the delta method of Stuart and Kendall (1977), Woodruff and Feldt derived the following estimate:

$$S_{ij} = 2\hat{\rho}_{ij}^2/9(\tilde{N} - 1)(1-\hat{\zeta}_i)^{1/3}\,(1-\hat{\zeta}_j)^{1/3} \quad . \tag{12}$$

As in the case of two coefficients, $\hat{\rho}_{ij}^2$ is the square of the sample correlation between the scores on tests $\underline{i}$ and $\underline{j}$. When the tests differ in length, then $\tilde{N} = N(\tilde{n}-1)/(\tilde{n}+1)$, where $\tilde{n}$ is the harmonic mean of all test lengths. On the assumption that the variables $1/(1-\hat{\zeta}_i)^{1/3}$ have a multivariate normal distribution, a matrix function of $\hat{\zeta}_i$, ρ, σ_i^2, σ_{ij}, and $\tilde{N}$ is shown to be distributed approximately as χ^2 with m-1 degrees of freedom. Woodruff and Feldt (in press) further show that an approximation of this function serves satisfactorily as a test statistic. It is

$$UX_1 = \sum_i^m [(1-\hat{\zeta}_1)^{-1/3} - \mu^*]^2 \;/\; (\bar{S}^2-\bar{C}) \;, \tag{13}$$

where $\bar{S}^2$ is the average of the variances S_i^2 (equation 11), $\bar{C}$ is the average of the covariances S_{ij} (equation 12), and μ^* is the average of the transformed coefficients, $1/(1-\hat{\zeta}_i)^{1/3}$. UX_1 is distributed approximately as chi-square with m-1 degrees of freedom when H_o is true.

This procedure may be illustrated by a four-test situation with N = 100 and the following data:

$$\left.\begin{array}{llll} \text{Test 1 : } n_1 = 100 & \hat{\zeta}_1 = .875 & (1-\hat{\zeta}_1)^{-1/3} = 2.0000 \\ \text{Test 2 : } n_2 = 75 & \hat{\zeta}_2 = .857 & (1-\hat{\zeta}_2)^{-1/3} = 1.9123 \\ \text{Test 3 : } n_3 = 50 & \hat{\zeta}_3 = .833 & (1-\hat{\zeta}_3)^{-1/3} = 1.8159 \\ \text{Test 4 : } n_4 = 25 & \hat{\zeta}_4 = .800 & (1-\hat{\zeta}_4)^{-1/3} = 1.7100 \end{array}\right\} \begin{array}{l} \tilde{n} = 48.0000 \\ \tilde{N} = 95.9184 \\ \mu^* = 1.85955 \end{array}$$

$$\text{Correlations: } \begin{bmatrix} 1.00 & .80 & .75 & .65 \\ & 1.00 & .70 & .60 \\ & & 1.00 & .55 \\ & & & 1.00 \end{bmatrix}$$

$$S_i^2 \text{ and } S_{ij}\text{: } \begin{bmatrix} .0093648 & .0057306 & .0047828 & .0033829 \\ & .0085615 & .0039836 & .0027561 \\ & & .0077201 & .0021991 \\ & & & .0068459 \end{bmatrix}$$

$\bar{S}^2 = .0081231$ $\quad \bar{C} = .0038059$ $\quad \bar{S}^2 - \bar{C} = .0043172$

$UX_1 = 10.836$ $\quad P\left[\chi^2_{(3)} > 10.836\right] = .013$

For this example, the more complex function for which UX_1 substitutes has the numerical value 11.118.

Analogous to the situation involving independent coefficients, follow-up tests of pairwise contrasts can be made via the t-test presented earlier for two coefficients.

Cruciality of the Statistical Assumptions

The most fundamental distributional assumption required by these inferential procedures is that the quantity $(1-\zeta)/(1-\hat{\zeta})$ be distributed

as F. As previously noted, this assumption will be met if the scores conform to the dictates of the two-way random effects model (type II ANOVA) with one observation per cell or the two-way mixed model (type III ANOVA) with one observation per cell and no interaction effect. These requirements will be met if the item or part scores are normally distributed with homogeneous error variances. However, these assumptions will almost surely be violated if each part of the instrument gives rise to a restricted range of scores. Therefore, the question arises how well the procedures may be expected to perform with actual data.

Feldt (1965), in deriving the F distribution for the transformed alpha coefficient, gives a detailed discussion of the assumptions required under the random effects model and how they might be violated with dichotomously scored items. He also reports on the results of a simulation study based on real test data with dichotomously scored items. The results indicate that the F distribution holds up well with such data.

In an experimental design context, Seeger and Gabrielsson (1968) simulate the distribution of mean square ratios under the mixed model ANOVA when applied to dichotomous data. They consider the situation involving several observations per cell and focus attention on the F ratio pertaining to treatment effects. Though this ratio is not the one used in reliability studies, their simulations offer further indirect evidence that $(1-\zeta)/(1-\hat{\zeta})$ is distributed approximately as F even if the items are dichotomously scored.

Inference for several alpha coefficients based on independent samples requires, in addition to distributional assumptions, that the sample sizes

be large enough to justify the asymptotic chi square distribution for the Hakstian/Whalen test statistic M and the Woodruff/Feldt statistic UX_2. Hakstian and Whalen used Monte Carlo methods to investigate the sampling distribution of test statistic M when computed from dichotomous part scores. Their results indicate good control of type I error rates with as few as twenty subjects per test, even for this gross departure from normality and homogeneity of variance.

If the same sample or matched samples are used for testing the equality of several alpha coefficients, two additional assumptions are required. The first is that the $1/(1-\hat{\zeta}_i)^{1/3}$ have a joint multivariate normal distribution. The second is that the correlations between total scores Y_i and Y_j be identical (homogeneous) for all pairs of tests. If the $(1-\zeta_i)/(1-\hat{\zeta}_i)$ have approximate F distributions, then the $1/(1-\hat{\zeta}_i)^{1/3}$ have marginal distributions approximately normal in form. Given these marginal normal distributions, it is reasonable to assume multivariate normality. However, multivariate normality does not automatically follow from the condition of marginal normality.

Woodruff and Feldt (in press) investigated the power and Type I error control of UX_1 using Monte Carlo methods. They found for a sample size as small as 50 and with moderately heterogeneous, positive inter-test correlations (range of ρ equal to .30), control of Type I error rates was quite good. However, these simulations were based on continuous, normally distributed scores. They did not provide evidence as to the cruciality of the normality assumption for total scores nor did they document the effects of dichotomous item scoring.

The results of subsequent Monte Carlo investigations of these issues are summarized in the tables which follow. In the first of these studies, dichotomous item scores were generated via a computer simulation technique described by Nitko (1968). Two true null hypotheses were considered. In the first, $\zeta_i = .80$ for each of four tests with 30 items in each test. In the second, $\zeta_i = .65$ for each of three tests with 30, 30, and 60 items, respectively. Each 30-item test exhibited a range of item difficulties from .30 to .80; the 60-item test had a range of item difficulties of .35 to .73. The item difficulty distribution for each test was unimodal and symmetrical around the value .55. The resultant distributions of total test scores were slightly skewed negatively ($\gamma_1 = -.13$) and platykurtic ($\gamma_2 = -.53$), generally similar to the distributions for many standardized tests. The inter-test correlations were homogeneous and equal to their shared reliability (.65 or .80).

For each null hypothesis, 2200 simulations of random sample data were produced, based on N = 50 and N = 100. Test UX_1 was performed on each replication, and the percent of test statistics exceeding the upper 10%, 5%, and 1% points of $\chi^2_{(m-1)}$ was tabulated. These data are summarized in Table 1.

It may be observed that the UX_1 test showed no gross effects from dichotomous item scoring. There is a tendency toward liberality if N=50 and a 10 or 5 percent level is employed, but this deviation from the nominal significance level would not disturb most researchers.

The second empirical study used actual test data--scores of Iowa students in grade 9 and 11 on various subtests of the Iowa Tests of

Table 1

Estimated Probability of Type I Error Based on 2200 Replications of the UX_1 Test: Simulated Dichotomous Item Scores

	ζ = .8	m = 4	n_i = 30	ζ = .65	m = 3	n_i = 60,30,30
	10%	5%	1%	10%	5%	1%
N=50	10.7	5.4	1.4	10.0	5.1	0.9
N=100	10.9	5.6	1.1	9.8	5.0	1.1

Educational Development, Form X-7. The subtests for grade 9 were selectively shortened by the deletion of items so that all tests had $\zeta = .75$. The subtests for grade 11 were differentially shortened so that all tests had $\zeta = .87$. From the pool of 16,443 records for grade 9 and 16,760 records for grade 11, 2,000 random samples of N=50 and 2,000 samples of N=100 were chosen by sampling examinees randomly with replacement. The UX_1 test was then executed on each examinee sample, using m=2, 3, 4, or 5 ITED subtests. Within the value of m=3 and 4, two groups of subtests were investigated. The first group exhibited less heterogeneity of inter-test correlations than did the second, but in both cases the null hypothesis with respect of ζ was true. The results of this study are summarized in Tables 2 and 3.

With actual test data the control of Type I error was not as tight as with simulated dichotomous item scores. The deviations from the nominal significance level were most pronounced with N=50, though not consistently in the direction of liberality. With m=3, for example, the deviations at 10% and 5% levels are positive for one group of tests and negative for the other. It must be borne in mind, of course, that the standard error of a percent in the vicinity of 10% equals about 0.67% with 2000 trials; near 5% the standard error equals about 0.49%.

A crude summary over the twelve situations for N=50 and N=100 gives rise to the following averages:

N=50	10.6%	5.5%	1.3%
N=100	9.9%	5.0%	1.1%

Table 2

Estimated Probability of type I Error Based on 2000 Replications of the UX_1 Test: Actual Dichotomous Item Scores (ζ = .75)

Sample Size	m=2 n_i=11,13 and n_i=11,16			m=5 n_i=11,13,16,21,21		
	10%	5%	1%	10%	5%	1%
N=50	10.4	5.6	1.2	11.9	6.0	1.1
N=100	9.8	4.7	0.9	10.3	5.2	0.8
	m=4 n_i=11,16,21,21		$.60<\rho_{xy}<.67$	m=4 n_i=11,13,16,21		$.55<\rho_{xy}<.65$
N=50	11.4	6.4	1.5	9.5	5.3	0.8
N=100	10.5	5.1	0.9	9.0	4.5	0.8
	m=3 n_i=11,16,21		$.60<\rho_{xy}<.65$	m=3 n_i=11,13,16		$.55<\rho_{xy}<.65$
N-50	9.9	4.9	1.3	10.4	5.6	0.9
N=100	9.1	4.0	0.6	10.4	4.1	1.1

Table 3

Estimated Probability of Type I Error Based on 2000 Replications of the UX_1 Tests: Actual Dichotomous Item Scores (ζ = .87)

Sample Size	m=2 n_i=24,34 and n_i=28,46			m=5 n_i=24,28,31,36,46		
	10%	5%	1%	10%	5%	1%
N=50	10.4	5.5	1.4	10.2	5.2	1.4
N=100	9.7	5.0	1.2	9.7	5.5	1.7
	m=4 n_i=24,36,46,28		.70<ρ_{xy}<.79	m=4 n_i=31,24,46,28		.63<ρ_{xy}<.79
N=50	12.4	6.9	1.7	10.3	4.7	1.2
N=100	10.8	5.8	1.5	9.8	5.5	1.4
	m=3 n_i=11,16,34		.60<ρ_{xy}<.65	m=3 n_i=31,24,46		.63<ρ_{xy}<.76
N=50	11.8	6.0	1.6	8.6	4.4	1.2
N=100	11.0	5.9	1.1	8.8	4.3	0.9

These means are very close to the averages for simulated data (Table 1). Together, they support the conclusion that the UX_1 test works quite well with N=100, but it errs on the side of liberality with N=50. The degree of liberality isn't great, and most researchers would probably be willing to accept a test that controls Type I error within one-half of one percent. But there is a need for an improved test for use with sample sizes of 50 or less. It is pertinent to note that almost all of the test instruments used in this study gave rise to negatively skewed, platykurtic score distributions. The γ_1 index of skewness ranged between -.597 and +.165, with eight of the ten indices negative. The γ_2 index of kurtosis ranged between -.015 and -.948. The average value of γ_2 for all ten tests (five in each of two grades) was -.676. Quite possibly this characteristic of the score distributions accounts for the liberality of the UX_1 test with the smaller sample size.

References

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56, 52-64.

Ebel, R. L. (1951). Estimation of reliability of ratings. Psychometrika, 16, 407-424.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. Psychometrika, 30, 357-370.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34, 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika, 45, 99-105.

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. Psychometrika, 41, 219-231.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart, and Winston.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Kendall, M. G. & Stuart, A. (1977). The Advanced Theory of Statistics (Vol. I, 4th ed.). London: Griffin.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika. 28, 221-238.

Marascuilo, L. A. (1966). Large-sample multiple comparisons. Psychological Bulletin, 65, 280-290.

Nitko, A. J. (1968). The power functions of some proposed test for the significance of coefficient alpha in the one-sample and two-sample cases. Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA.

Oaster, R. R. F. (1984). Even numbered alternative scale reliability. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Paulson, E. (1942). An approximate normalization of the analysis of variance distribution. Annals of Mathematical Statistics, 13, 233-235.

Pitman, E. J. G. (1939). A note on normal correlation. Biometrika, 31, 9-12.

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

Seeger, P., Gabrielsson, A. (1968). Applicability of the Cochran Q tests and F test for statistical analysis of dichotomous data for dependent samples. Psychological Bulletin, 69, 269-277.

Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. Proceedings of the National Academy of Sciences, USA, 17, 684.

Woodruff, D. J., & Feldt, L. S. (in press). Tests for equality of several alpha coefficients when their sample estimates are dependent. Psychometrika.